

Breaking Barriers in Data Access & Sharing: A Bottom-up View

Françoise Genova



“Riding the wave” report: Recom 1

1. Develop an international framework for a Collaborative Data Infrastructure

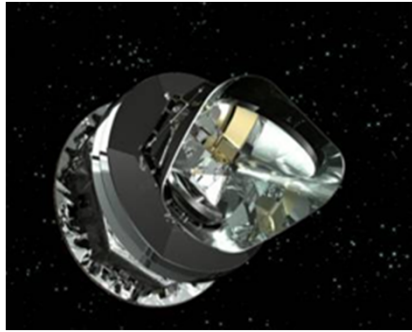
The scientific community is supported to provide its data and metadata for re-use.

Every funded science project includes a fixed budget percentage for compulsory conservation and distribution of data, spent depending on the project context.

Data form an infrastructure

Astronomy infrastructures





Why share and reuse data

- In disciplines which use large “physical” infrastructures, it can be difficult to understand that data can also be an infrastructure... There should be a political incentive to reuse data so that the use of taxpayer money is optimized, but the data infrastructure had to demonstrate that it is useful
- In other disciplines such as human sciences data can be seen as the core infrastructure (cf ESFRI in Europe)
- Political decisions are not enough
- The science community has to agree fully

The role of science communities

- Science community has to engage itself at several levels
 - For all projects, part of project budget has to go to data sharing
 - Committees members have to be convinced, since this is in competition with support to new science programmes or new instruments
 - Active participation of the community is required
 - Strong disciplinary “pillars” are needed to do interdisciplinary science and people are needed to build them
- Sharing data is not only piling up data in a repository

Enabling science is a major aim

- Sharing and re-using data
Data + metadata + tools
- Has to be done in a way so that it is useful for science to get communities' approval and participation
- If not done the right way potential users will rapidly be discouraged with strong consequences on the whole process

Work closely with real users and build according to their requirements.

With a science driven data infrastructure

Dramatic progress in the efficiency of the scientific process, and rapid advances in our understanding of our complex world, enabling the best brains to thrive wherever they are.

A powerful tool towards global integration
of the scientific community
An important aim e.g. for Europe

The “it’s my data” syndrome

- A blunt general political decision to open data is not enough since it has to be felt as legitimate by people who produce the data
- Have it in mind when proposing directives
- Envisage practical policies such as protections adapted to the case or as limited “proprietary periods” so that data producers have some time to use it themselves before sharing it
- Keep track of who has produced the data all along the data life (“provenance”)

The career bottleneck

Propose reliable metrics to assess the quality and impact of datasets. All agencies should recognise high quality data publication in career advancement.

- We have heard that so many times... How to make it happen in this era of dominant h- and other indexes ?
- Two aspects
 - Data publication by the scientists who produce them should be recognized as one of their impacts
 - Data curation by data scientists

Create the discipline of data scientist, to ensure curation and quality in all aspects of the system.

- A profile for scientists: data curation is a high impact scientific task when the data is widely used and scientific skills are required to do it well
- Another useful profile which should also get recognized is an evolution of librarians' profile, requiring additional disciplinary knowledge and working in (or in very close connection with) disciplinary data teams

Networking and Interoperability

- Data access is a first step, but the user has to find his/her own way in the “jungle” of isolated on-line data resources
- Networking resources through web links is a powerful improvement : users can navigate from one resource to another
- Interoperability allows data discovery, retrieval and usage across the borders of the individual data resources

Interoperability within and across disciplines

- Global interoperability is a very complex problem
- Diversity is a keyword of the data infrastructure
- Data access probably not fully done anywhere but disciplines are at very different levels of preparedness
 - Some already have a disciplinary data system (or rather “ecosystem”), and some have developed interoperability
 - Other have not really begun
 - Many are in between

Accommodate diversity

- The Collaborative Data Infrastructure must accommodate this diversity, with some data systems built on generic solutions and others which are already well developed
- Do not kill what is working, but build upon it
- The proposed CDI framework must be acceptable for data providers (small overhead when joining it)
- A system of systems with a “light” interoperability structure on top of existing data holdings/data systems?

The technology, stupid!

- Technology is of course one of the cornerstones, together with data curation and users
- This is true at all steps, from data provision to the building of an interoperability framework
- Adequate knowledge and usage of technologies is a key of sustainability (cf CDS 40 years – how many technological evolutions/revolutions during the last 40 years?)
- Another profile in the teams at all levels, particularly mandatory in the discussion of the interoperability framework
- Beware of the buzz, look for “sustainable enough” solutions

Too complex to work

Do not aim for a single top down system

Ensure effective governance and maintenance system

- Decompose the very complex general problem into manageable building blocks (and identify dependencies)
- Identify the critical building blocks
 - Persistent Identifiers are critical for many topics (e.g., they can play an important role in the linkage of data and publications). What are the key requirements? Lots of studies already.
 - Registry is the key to data discovery and retrieval
 - Semantics at disciplinary and interdisciplinary levels – but how to focus the work in that domain?
 - ...

Among the other risks

The infrastructure is not used

Work closely with real users and build according to their requirements

Lack of willingness to co-operate across disciplines/ funders/ nations

Apply subsidiarity principle so we do not step on researchers' toes

Take advantage of growing need of integration: within and across disciplines

Provide “forums” to define strategies at disciplinary and cross-disciplinary levels for metadata definition.

- One key role of the RDA should be to define the basic interoperability building blocks
- Ensure that a sufficiently varied set of disciplines participate, and include data providers and users as well as “technologists”
- Lots has been done by disciplines and generic system providers
 - Identify commonalities, discuss lessons learnt, identify what works
- Have in mind real examples of cross-discipline needs

Measurement of success

- Time needed to agree internationally on standards. Set focussed aims and milestones.
- Take-up by data providers and usage by the science community and others
- System open to other usages

- Set up the right structure, procedures and aims
- And gather the right people to do the job
- Let's do it!